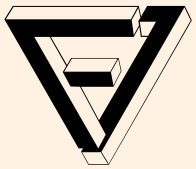


Improving the Reliability of the Plagiarism Detection System

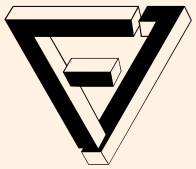
Jan Kasprzak and Michal Brandejs

Faculty of Informatics, Masaryk University
Brno, Czech Republic

Czech Republic

Brno

... birthplace of Kurt Gödel
(theorems about incompleteness)

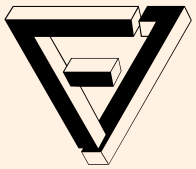


From Where?

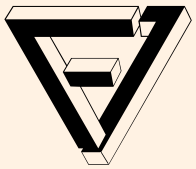
Masaryk University

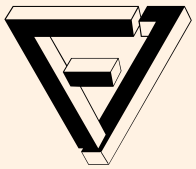
About 40,000 students in 9 faculties.
Named after the first president of Czechoslovakia.

Faculty of Informatics



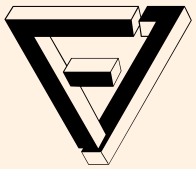
Agenda





The Information System

- **Masaryk University Information System**
- `http://is.muni.cz/?lang=en`
- 30,000 unique users daily
- 2,000,000 HTTP requests daily on average
  - Monday, Sep 20 record: > 3,000,000 requests
- 20,000,000 documents in storage:
  - theses,
  - study materials,
  - seminar works,
  - discussion forum posts,
  - etc.

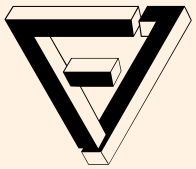


The Theses Archive

- **Czech National Archive of Graduate Theses**
- `http://theses.cz/`
- Theses metadata and full texts

So our motivation is:

We need a working plagiarism detection system.

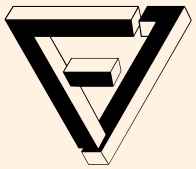


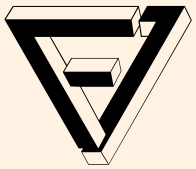
The Theses Archive

- **Czech National Archive of Graduate Theses**
- `http://theses.cz/`
- Theses metadata and full texts

So our motivation is:

We need a working plagiarism detection system.





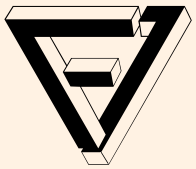
Our PAN'09 Sysem

- Starting point for PAN 2010.
- External plagiarism only.

2009	Recall	Prec.	Gran.	Overall
1. Grozea	0.6585	0.7418	1.0038	0.6957
2. Kasprzak	0.6967	0.5573	1.0228	0.6093

Interpretation

For PAN 2010, focus on **precision** and **granularity**.



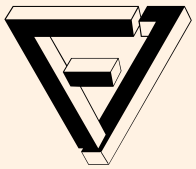
Our PAN'09 Sysem

- Starting point for PAN 2010.
- External plagiarism only.

2009	Recall	Prec.	Gran.	Overall
1. Grozea	0.6585	0.7418	1.0038	0.6957
2. Kasprzak	0.6967	0.5573	1.0228	0.6093

Interpretation

For PAN 2010, focus on **precision** and **granularity**.



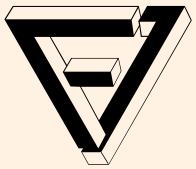
Our PAN'09 Sysem

- Starting point for PAN 2010.
- External plagiarism only.

2009	Recall	Prec.	Gran.	Overall
1. Grozea	0.6585	0.7418	1.0038	0.6957
2. Kasprzak	0.6967	0.5573	1.0228	0.6093

Interpretation

For PAN 2010, focus on **precision** and **granularity**.

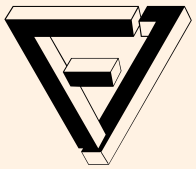


PAN'09 System Structure

A very brief outline:

1. Tokenization of source documents
2. Chunks, their fingerprints and position data
3. Inverted index
4. Suspicious docs: tokenization, chunks, fingerprints
5. Lookup in the inverted index
6. Valid intervals of common chunks
   - Positions in both suspicous and source document should not be too far apart.
7. Postprocessing
   - Removing overlaps etc.

See our paper for PAN'09.

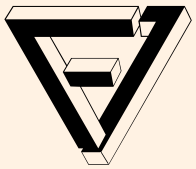


Training data: PAN-PC-09

- United external and intrinsic data
  - to get an estimate for PAN 2010

	Recall	Prec.	Gran.	Overall
PAN'09	0.5255	0.4858	1.0480	0.4882

This is the baseline of our PAN 2010 work.

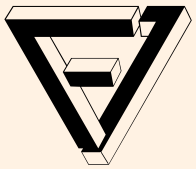


Training data: PAN-PC-09

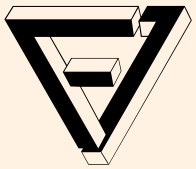
- United external and intrinsic data
  - to get an estimate for PAN 2010

	Recall	Prec.	Gran.	Overall
PAN'09	0.5255	0.4858	1.0480	0.4882

This is the baseline of our PAN 2010 work.



External Detector Improvements

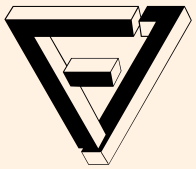


Overlapping Detections

- PAN'09: keep the longer one
- Idea: if both are *short*, remove them both!
- Implementation: *short* is < 600 characters

	Recall	Prec.	Gran.	Overall
Baseline	0.5255	0.4858	1.0480	0.4882
Overlaps	0.5252	0.4941	1.0465	0.4929

Possible reason: common phrases or constructs.

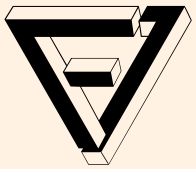


Overlapping Detections

- PAN'09: keep the longer one
- Idea: if both are *short*, remove them both!
- Implementation: *short* is < 600 characters

	Recall	Prec.	Gran.	Overall
Baseline	0.5255	0.4858	1.0480	0.4882
Overlaps	0.5252	0.4941	1.0465	0.4929

Possible reason: common phrases or constructs.

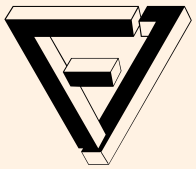


Adjacent Detections

- Improve the granularity.
- Join adjacent detections
  - from the same source document.
- Maximum gap should depend on the detections size.
- Algorithm:
  - Gap < 600 chars: **merge**
  - Gap < 4000 characters and smaller than half of average length of both detections: **merge**
  - Otherwise: **keep separated**.

	Recall	Prec.	Gran.	Overall
Overlaps	0.5252	0.4941	1.0465	0.4929
Merge	0.5256	0.5302	1.0233	0.5192

Improved both precision and granularity.

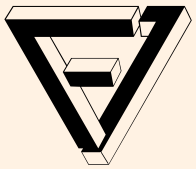


Adjacent Detections

- Improve the granularity.
- Join adjacent detections
  - from the same source document.
- Maximum gap should depend on the detections size.
- Algorithm:
  - Gap < 600 chars: **merge**
  - Gap < 4000 characters and smaller than half of average length of both detections: **merge**
  - Otherwise: **keep separated**.

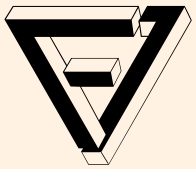
	Recall	Prec.	Gran.	Overall
Overlaps	0.5252	0.4941	1.0465	0.4929
Merge	0.5256	0.5302	1.0233	0.5192

Improved both precision and granularity.



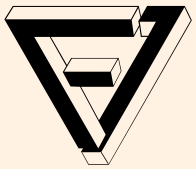
False Positives

- In PAN'09: tables of contents, tables of references, etc.
- Ideas:
  - Structure of text (line wrapping, etc.).
  - Non-letter characters (see Stamatatos, 2009).
- Exclude passages with low ratio of letters.



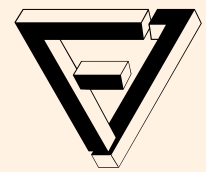
False Positives

- In PAN'09: tables of contents, tables of references, etc.
- Ideas:
  - Structure of text (line wrapping, etc.).
  - Non-letter characters (see Stamatatos, 2009).
- Exclude passages with low ratio of letters.

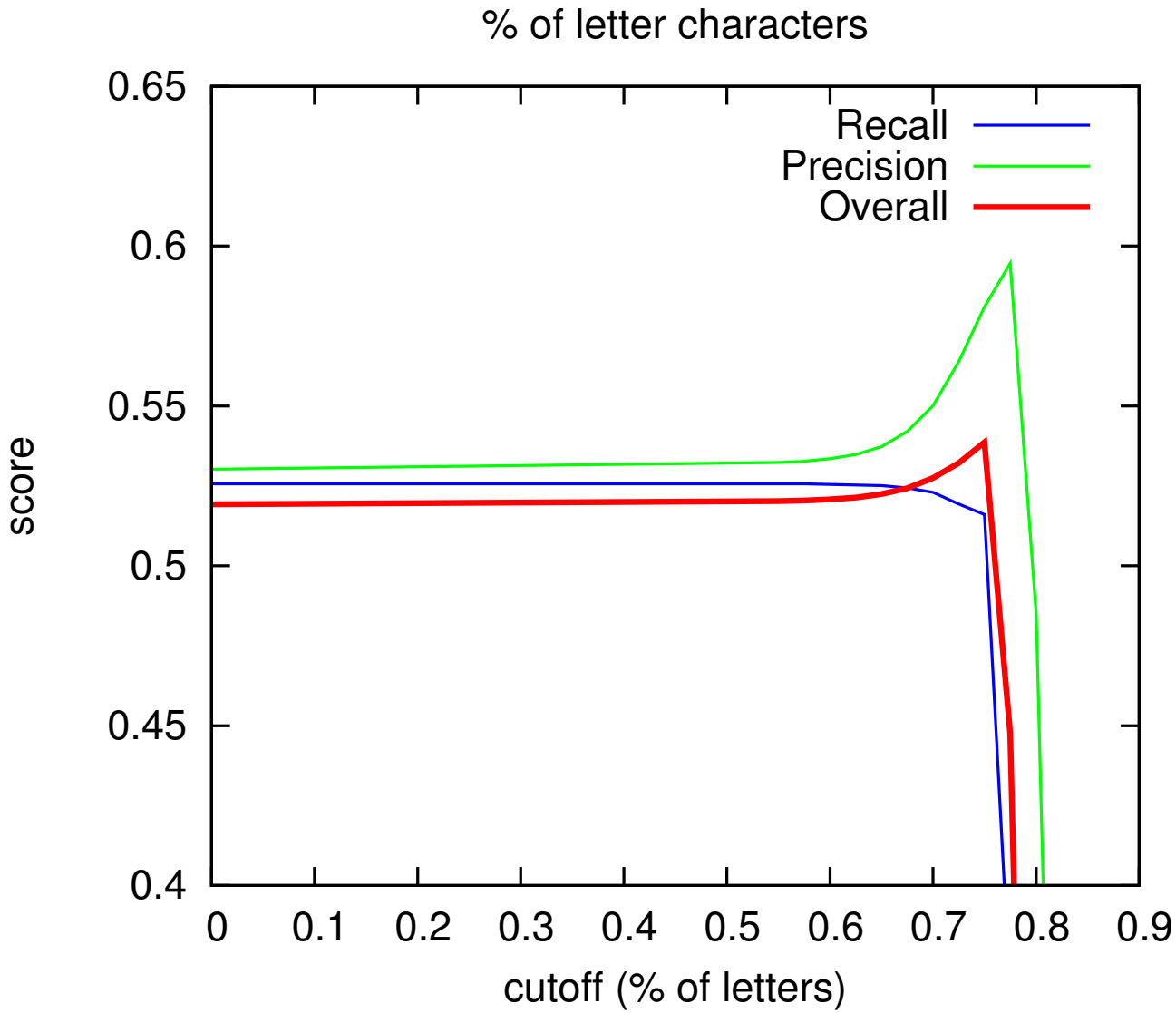


False Positives

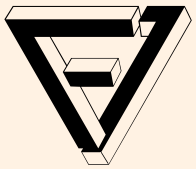
- In PAN'09: tables of contents, tables of references, etc.
- Ideas:
  - Structure of text (line wrapping, etc.).
  - Non-letter characters (see Stamatatos, 2009).
- Exclude passages with low ratio of letters.



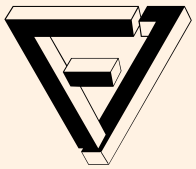
Letter Characters Ratio



The final threshold used was **0.675**.

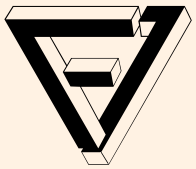


Cross-language Plagiarism Detector



Naive Approach

- Source documents: English, German, and Spanish only
- Suspicious documents: English only
- Use the machine translation
  - ... and hope the results will be similar enough
- Implemented after the last deadline extension

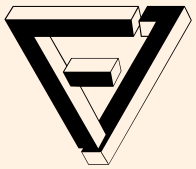


Language Detection

- `Text::Language::Guess` Perl module
- Stop-words based classification
- Many misdetections
  - e.g. PAN-PC-09 document 112 detected as French
- Non-english results were checked by hand
- Ready-to-use, fast enough

Suggestion

A classifier based on n-gram character profiles would probably be better.

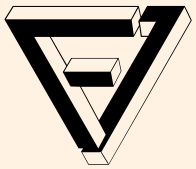


Language Detection

- `Text::Language::Guess` Perl module
- Stop-words based classification
- Many misdetections
  - e.g. PAN-PC-09 document 112 detected as French
- Non-english results were checked by hand
- Ready-to-use, fast enough

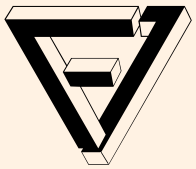
Suggestion

A classifier based on n-gram character profiles would probably be better.



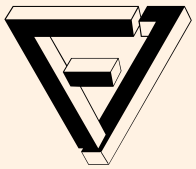
Machine Translators

- Yahoo! Babelfish
  - Long timeouts
  - Sometimes did not respond at all
  - Does not keep formatting
- Google Translate
  - Keeps line breaks!
  - Sometimes truncates the output



Machine Translators

- Yahoo! Babelfish
  - Long timeouts
  - Sometimes did not respond at all
  - Does not keep formatting
- Google Translate
  - Keeps line breaks!
  - Sometimes truncates the output

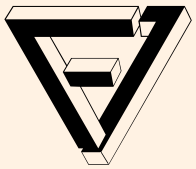


Machine Translation

- Google Translate
- 15–22 KB requests
- Split at paragraph boundary, if possible
- Otherwise, split at line breaks

Data for translator:

- 2562 parts for PAN-PC-09 Spanish
- PAN-PC-09 German omitted
- 4887 parts for PAN-PC-10 German
- 2562 parts for PAN-PC-10 Spanish

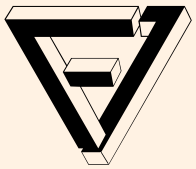


Problematic Sentences

PAN-PC-10 document 6696 line 6256

unser Los. Und ich bin ja auch glücklich, wenn ich nur weiß, daß Moina sich vergnügt.< Sie

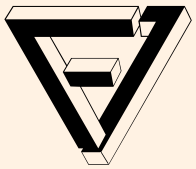
- Processing always stopped after the word *Moina*.
- Even in single-line request.
- Problematic sentences/blocks replaced by empty lines.

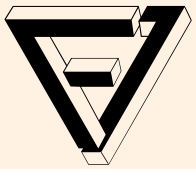


Cross-Language Detection Results

	Recall	Prec.	Gran.	Overall
< 0.675	0.5244	0.5420	1.0233	0.5243
Spanish	0.5386	0.5476	1.0236	0.5340

Bigger improvement expected for competition corpus (German as well).



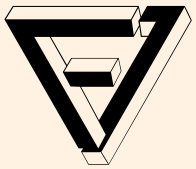


Intrinsic Detector Outline

Stamatatos, 2009:

- Partly overlapping windows
- Character trigram frequencies
- Style change function *sc(win)*
  - Window versus the whole document
- Higher *sc(win)* marks plagiarized passage

Refer to the original article for details.



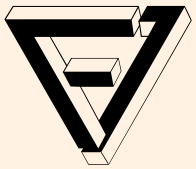
Our Improvements

Reimplementation of Stamatatos' approach

Could not reproduce the score of 0.2462

- our was about 0.172 at most

- Different means of determining the plagiarized passage
- Different window endpoints



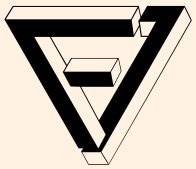
Our Improvements

Reimplementation of Stamatatos' approach

Could not reproduce the score of 0.2462

- our was about 0.172 at most

- Different means of determining the plagiarized passage
- Different window endpoints



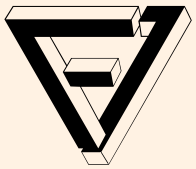
Our Improvements

Reimplementation of Stamatatos' approach

Could not reproduce the score of 0.2462

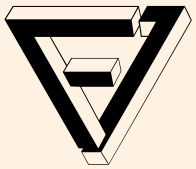
- our was about 0.172 at most

- Different means of determining the plagiarized passage
- Different window endpoints



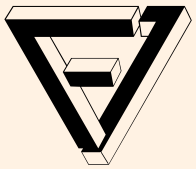
Smoothed Style-Change Function

- Motivation: find "generally high" areas
- Gaussian-weighted averaging
- Two averaged functions: $\sigma = 1$, $\sigma = 10$.
- Plagiarized passage boundary:
  - Smoothed style change functions intersect each other,
  - the neighbouring local minima/maxima are low/high enough

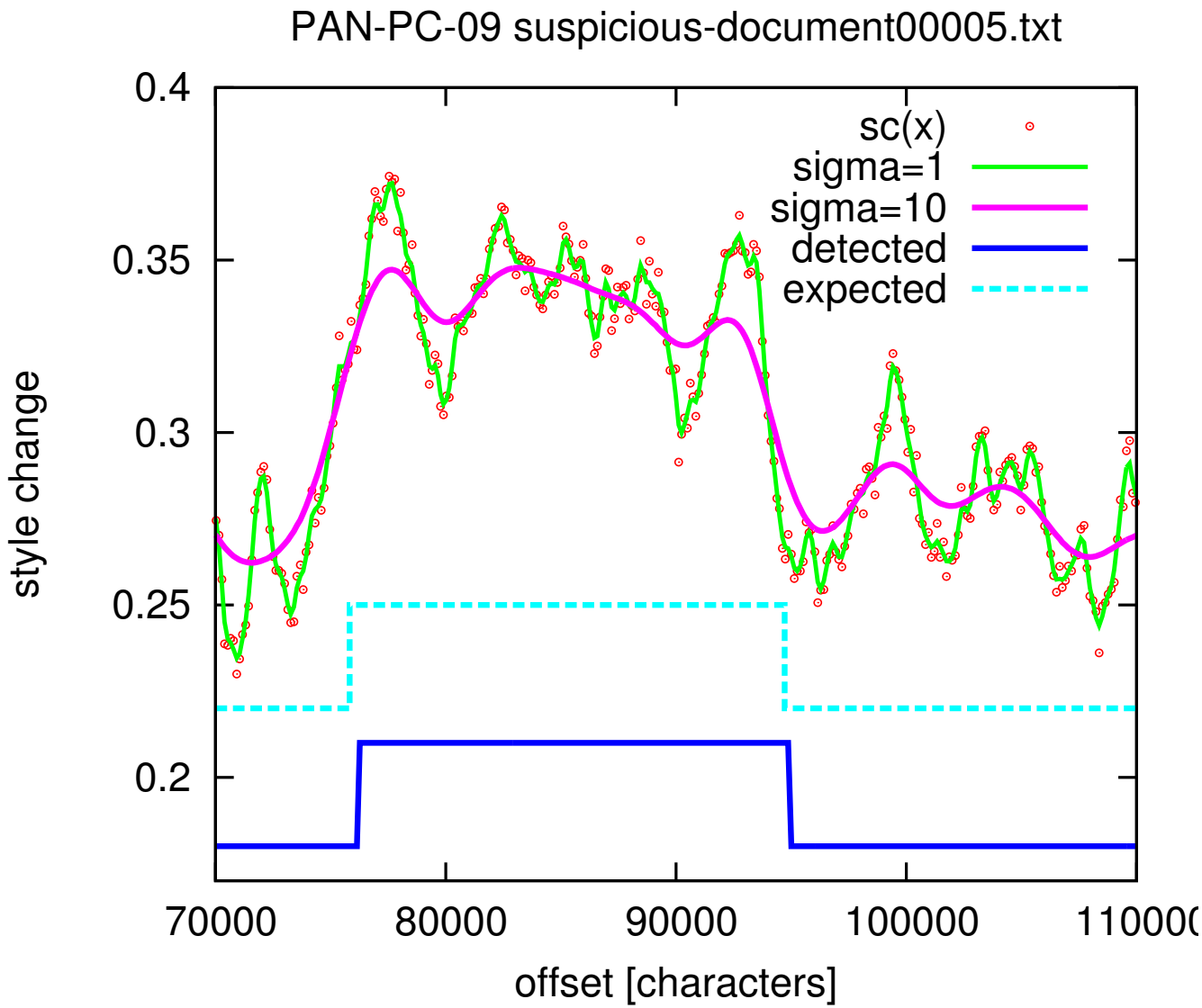


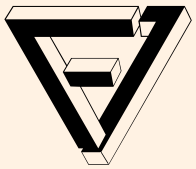
Smoothed Style-Change Function

- Motivation: find "generally high" areas
- Gaussian-weighted averaging
- Two averaged functions: $\sigma = 1$, $\sigma = 10$.
- Plagiarized passage boundary:
  - Smoothed style change functions intersect each other,
  - the neighbouring local minima/maxima are low/high enough



Intrinsic Detector: Example





Window Sizing

- Not in terms of character count,
- but in terms of trigram count.

Possible explanation:

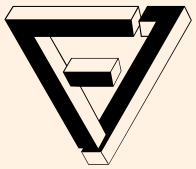
- *sc(win)* is not as stable as stated.

	Recall	Prec.	Gran.	Overall
Stamatatos	0.4607	0.2321	1.3839	0.2462
Kasprzak	0.2627	0.2969	1.072	0.2562

Future work

Different window-to-document distance.
E.g.: Out-of-place n-gram distance.

We did not use the intrinsic detector after all.



Window Sizing

- Not in terms of character count,
- but in terms of trigram count.

Possible explanation:

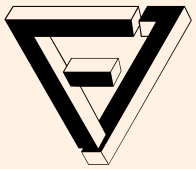
- *sc(win)* is not as stable as stated.

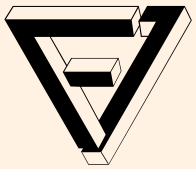
	Recall	Prec.	Gran.	Overall
Stamatatos	0.4607	0.2321	1.3839	0.2462
Kasprzak	0.2627	0.2969	1.072	0.2562

Future work

Different window-to-document distance.
E.g.: Out-of-place n-gram distance.

We did not use the intrinsic detector after all.





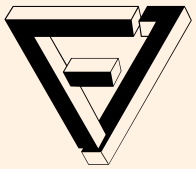
2010 Improvements on PAN-PC-09

Recapitulation (2009 data):

PAN-PC-09	Recall	Prec.	Gran.	Overall
Baseline	0.5255	0.4858	1.0480	0.4882
Overlaps	0.5252	0.4941	1.0465	0.4929
Merge	0.5256	0.5302	1.0233	0.5192
> 0.675	0.5244	0.5240	1.0233	0.5243
Spanish	0.5386	0.5476	1.0236	0.5340

Summary

- Improved precision and granularity.
- Overall improvement not very significant.

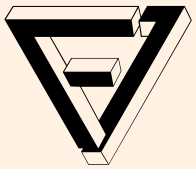


PAN-PC-10 Performance

	Recall	Prec.	Gran.	Overall
PAN-PC-09	0.5386	0.5476	1.0236	0.5340
PAN-PC-10	0.6915	0.9405	1.0004	0.7968

- Unexpectedly high precision
- Recall close to theoretical maximum (w/o intrinsic)

But how did we get there?

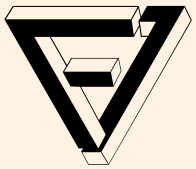


PAN-PC-10 Performance

	Recall	Prec.	Gran.	Overall
PAN-PC-09	0.5386	0.5476	1.0236	0.5340
PAN-PC-10	0.6915	0.9405	1.0004	0.7968

- Unexpectedly high precision
- Recall close to theoretical maximum (w/o intrinsic)

But how did we get there?

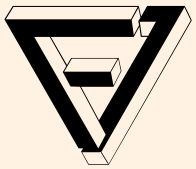


2010 Improvements on PAN-PC-10

PAN-PC-10	Recall	Prec.	Gran.	Overall
Baseline	0.6318	0.9140	1.0072	0.7432
Overlaps	0.6317	0.9147	1.0072	0.7435
Merge	0.6309	0.9243	1.0005	0.7497
> 0.675	0.6305	0.9264	1.0005	0.7500
ES + DE	0.6915	0.9405	1.0004	0.7968

Discussion

- Last year's SW would have also won
- Improvements even less significant on 2010 data
- Except translations
  - about 5 % on the overall score

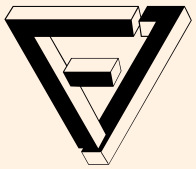


2010 Improvements on PAN-PC-10

PAN-PC-10	Recall	Prec.	Gran.	Overall
Baseline	0.6318	0.9140	1.0072	0.7432
Overlaps	0.6317	0.9147	1.0072	0.7435
Merge	0.6309	0.9243	1.0005	0.7497
> 0.675	0.6305	0.9264	1.0005	0.7500
ES + DE	0.6915	0.9405	1.0004	0.7968

Discussion

- Last year's SW would have also won
- Improvements even less significant on 2010 data
- Except translations
  - about 5 % on the overall score

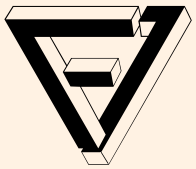


2010 Improvements on PAN-PC-10

PAN-PC-10	Recall	Prec.	Gran.	Overall
Baseline	0.6318	0.9140	1.0072	0.7432
Overlaps	0.6317	0.9147	1.0072	0.7435
Merge	0.6309	0.9243	1.0005	0.7497
> 0.675	0.6305	0.9264	1.0005	0.7500
ES + DE	0.6915	0.9405	1.0004	0.7968

Discussion

- Last year's SW would have also won
- Improvements even less significant on 2010 data
- Except translations
  - about 5 % on the overall score

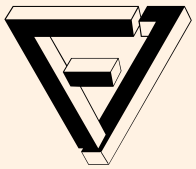


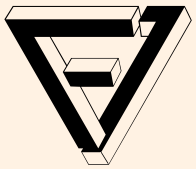
2010 Improvements on PAN-PC-10

PAN-PC-10	Recall	Prec.	Gran.	Overall
Baseline	0.6318	0.9140	1.0072	0.7432
Overlaps	0.6317	0.9147	1.0072	0.7435
Merge	0.6309	0.9243	1.0005	0.7497
> 0.675	0.6305	0.9264	1.0005	0.7500
ES + DE	0.6915	0.9405	1.0004	0.7968

Discussion

- Last year's SW would have also won
- Improvements even less significant on 2010 data
- Except translations
  - about 5 % on the overall score

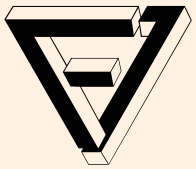




Lessons Learned

- The chunking algorithm works.
- The implementation does matter.
  - reading papers is not enough
  - see the intrinsic detector differences
- Some improvements unusable in real world.
  - e.g. machine translations
- PAN-PC-10 structure is substantially different to PAN-PC-09.

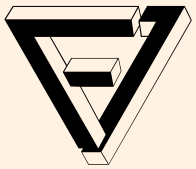
About our participation in PAN 2010 ...



Lessons Learned

- The chunking algorithm works.
- The implementation does matter.
  - reading papers is not enough
  - see the intrinsic detector differences
- Some improvements unusable in real world.
  - e.g. machine translations
- PAN-PC-10 structure is substantially different to PAN-PC-09.

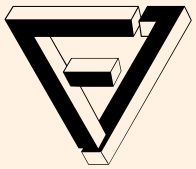
About our participation in PAN 2010 ...



Lessons Learned

- The chunking algorithm works.
- The implementation does matter.
  - reading papers is not enough
  - see the intrinsic detector differences
- Some improvements unusable in real world.
  - e.g. machine translations
- PAN-PC-10 structure is substantially different to PAN-PC-09.

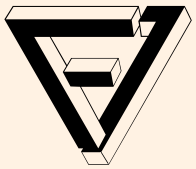
About our participation in PAN 2010 ...



Lessons Learned

- The chunking algorithm works.
- The implementation does matter.
  - reading papers is not enough
  - see the intrinsic detector differences
- Some improvements unusable in real world.
  - e.g. machine translations
- PAN-PC-10 structure is substantially different to PAN-PC-09.

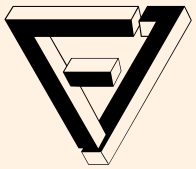
About our participation in PAN 2010 ...



Lessons Learned

- The chunking algorithm works.
- The implementation does matter.
  - reading papers is not enough
  - see the intrinsic detector differences
- Some improvements unusable in real world.
  - e.g. machine translations
- PAN-PC-10 structure is substantially different to PAN-PC-09.

About our participation in PAN 2010 ...



Our system in PAN 2010

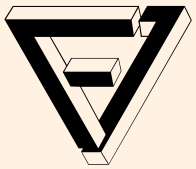
Does our system work?

Yes! We have got the first place in PAN 2010. Also:

- production use in `is.muni.cz` and `theses.cz`
- 2,000,000 of documents
- cluster-based implementation

Is it science?

Most probably not.
Ad-hoc improvements too tailored to the PAN-PC-09 structure.



Our system in PAN 2010

Does our system work?

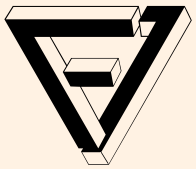
Yes! We have got the first place in PAN 2010. Also:

- production use in `is.muni.cz` and `theses.cz`
- 2,000,000 of documents
- cluster-based implementation

Is it science?

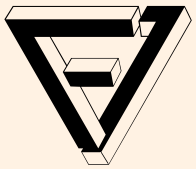
Most probably not.

Ad-hoc improvements too tailored to the PAN-PC-09 structure.



Questions?

Thanks for your attention!



System Scalability

- SGI Altix XE
  - Xeon E5472, 3.0 GHz, 8 threads total
  - 64 GB RAM
  - used also during PAN'09
- HP DL585 G6
  - Opteron 8439 SE, 2.8 GHz, 24 cores total
  - 128 GB RAM

Task	8 (SGI)	24 (HP)	speedup
Inv. index	1:06:02	0:12:41	520 %
Chunk pairs	2:07:25	0:20:44	615 %
Postproc.	0:09:22	0:03:17	285 %
Total	**3:22:55**	**0:36:42**	**553 %**